VEROVATNOĆA, ANALIZA PODATAKA, PARADOKSI I BAYES

10. 07. 2015.

KO GOD MISLIO DA ZNA…

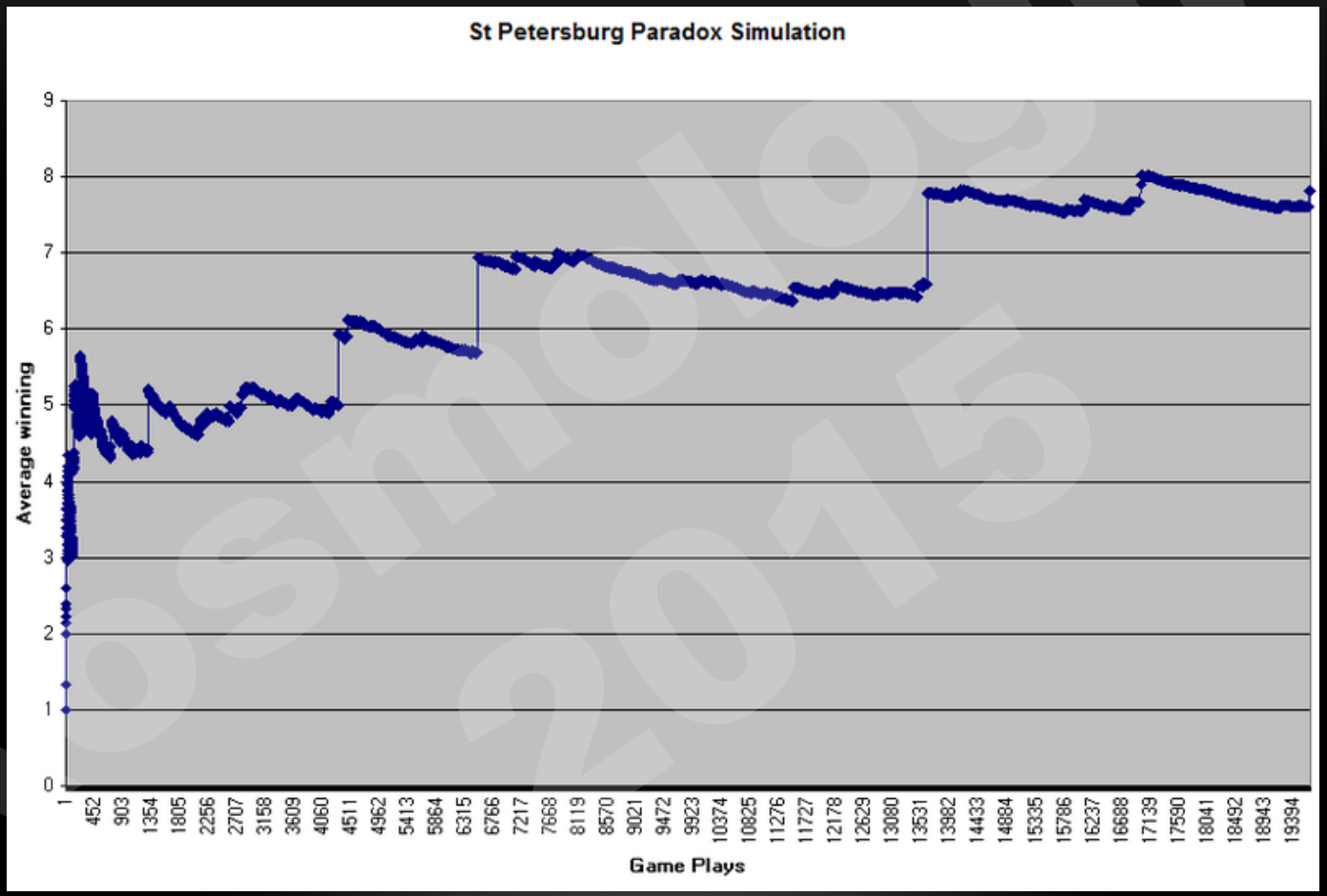
- …često se vara!
- Brojni paradoksi
  - Paskalova opklada
  - Hempelov paradoks (“bele vrane”)
  - Sankt-Peterburški paradoks
  - Argument Sudnjeg dana (Doomsday Argument)
- **Dve velike škole**:
  - Frekventistička interpretacija verovatnoće
  - Bajesovska (“subjektivistička”) interpretacija verovatnoće
- Stvar ima ogromne praktične posledice!

ST. PETERSBURŠKI PARADOKS

- Bernoulli, N. 1713, vs. Bernoulli, D. 1738
- „Ulazak“ košta X evra, dobitaj počinje sa 1e, a udvostručava se svaki put kad padne glava; kad prvi put padne pismo, igra se završava
- Ako se baca *k* puta, igrač dobija 2^{k-1} evra
- Koliko X bi bilo fer platiti? Koliko bi kazino trebalo da traži?

$$E = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 4 + \frac{1}{16} \cdot 8 + \cdots$$

- Teorija očekivane korisnosti, odbacivanje očekivanja, težine verovatnoća...
- Motivacija za veliki rad u ekonomiji, psihologiji, primenenjenoj matematici


St Petersburg Paradox Simulation
Average winning
0
1
2
3
4
5
6
7
8
9
1
452
903
1354
1805
2256
2707
3158
3609
4060
4511
4962
5413
5864
6315
6766
7217
7668
8119
8570
9021
9472
9923
10374
10825
11276
11727
12178
12629
13080
13531
13982
14433
14884
15335
15786
16237
16688
17139
17590
18041
18492
18943
19394
Game Plays

FREKVENTISTI VS. BAYES-OVCI

- Verovatnoća (A) = objektivna relativna frekvencija dešavanja A.
- Parametri su fiksirane nepoznate **konstante**, tako da ne možemo pisati, na primer, $P(\theta=0.5|D)$.
- Estimatori treba da budu zadovoljavajući kada se usrednje preko mnogih pokušaja.
- *Slučajne promenljive*

- Verovatnoća (A) = stepen poverenja da će se dogoditi A (u svetlu svih neodređenosti).
- Možemo pisati P(bilo šta|D)!
- Estimatori treba da budu zadovoljavajući *za skup podataka na raspolaganju!*
- *Neizvesne* (engl. *uncertain*) *promenljive*

PROBLEMI FREKVENTISTIČKE INTERPRETACIJE

- Verovatnoća pojedinačnih slučajeva: smemo li da kažemo npr.
  - Sutra će verovatno padati kiša.
  - Radikali će najverovatnije pobediti na sledećim izborima.
  - Pera će verovatno doktorirati do kraja godine.
  - Verovatnoća otkrića kvazara na z > 6 je jako mala.

- **Definicija** verovatnoće
  - Definicija *limesa* zahteva **beskonačan** niz, što ne postoji u fizičkom svetu.
  - Kad je nešto "dovoljno" slučajno?

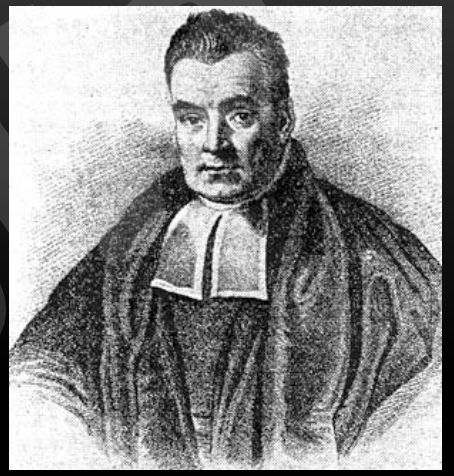
- „The dominant school in statistics since the beginning of last century is based on a quite unnatural approach to probability, in contrast to that of the founding fathers (Poisson, Bernoulli, Bayes, Laplace, Gauss, etc.). In this approach (frequentism) there is no room for the concept of probability of causes, probability of hypotheses, probability of the values of physical quantities, and so on. Problems in the probability of causes have been replaced by the machinery of the hypothesis tests. But people think naturally in terms of probability of causes, and the mismatch between natural thinking and standard education in statistics leads to the troubles discussed above.“

G. D‘Agostini (2004)

PRAVILO ZAUSTAVLJANJA?

- Šta nas iznenađuje – šta zahteva objašnjenje?
- Zaključci koji se učine iz podataka treba da zavise samo od sakupljenih podataka, ne od razloga zašto je baš ta količina podataka sakupljena.
- Ako pogledate podatke da biste odlučili kada da prestanete sa eksperimentom, to ne sme da promeni bilo koji zaključak koji ćete izvući!
- Klasični pristup analizi podataka često je u sukobu sa pravilom zaustavljanja.

BAYES-OVSKI PRISTUP



- Postoji mnogo objašnjenja bilo kog fenomena.
- Formulišem svako objašnjenje u formi hipoteze.
- Imam određeni razlog da dodelim *a priori* verovatnoću svakoj od mogućih hipoteza (“prior probability”)
- Nakon što se upoznam sa evidencijom, mogu da izračunam *a posteriori* raspodelu korišćenjem Bajesove formule.
- Tada mogu identifikovati objašnjenje sa hipotezom najveće aposteriorne verovatnoće.
- Ovo je **bajesovsko zaključivanje** (*the Bayesian inference*).

“NE MOŽ’ BIT’ PROSTIJI” PRIMER

- Test na AIDS savršeno tačno daje pozitivan nalaz ako je osoba inficirana, $P(+|inf.)=1$, ali ima malu verovatnoću od 0,2% za pozitivni nalaz ako je testirana osoba neinficirana: $P(+|\neg inf.)=0,002$.
- Ako dobijete pozitivan nalaz, sa kojim stepenom poverenja treba da verujete da ste stvarno zaraženi? $P(inf|+) = ?$
- (Tragična) ironija je da će i lekari skoro uvek reći 99,8%, što je...
- ...**POTPUNO POGREŠNO!!!** U gornjoj postavci nema dovoljno informacija da bi se smisleno odgovorilo na pitanje!
- Moguća dodatna (“apriorna”) informacija: gde se vrši testiranje?
- Ako se testirate u Evropi, verovatnoća da ste zaraženi iako imate pozitivan nalaz je svega oko 33%!!!

OSTATAK PRIČE

- Testiranje hipoteza – Bayesovski pristup
- Testiranje hipoteza – klasični (frekventistički) pristup
- Šta nije u redu sa klasičnim pristupom?
- Šta nije u redu sa Bayesovskim pristupom? (*Doomsday Argument*)

BAYES-OVO PRAVILO

Posteriorna verovatnoća

Uslovna verovatnoća

Apriorna verovatnoća

$$p(h \mid d) = \frac{p(d \mid h)\, p(h)}{\sum_{h' \in H} p(d \mid h')\, p(h')}$$

Sumiranje po prostoru hipoteza

POREKLO BAYES-OVOG PRAVILA

- Jednostavna posledica korišćenja verovatnoća da bi se predstavio stepen verovanja (poverenja)
- Za bilo koje dve slučajne promenljive:

$$p(A \& B) = p(A)\,p(B \mid A)$$

$$p(A \& B) = p(B)\,p(A \mid B)$$

$$p(B)\,p(A \mid B) = p(A)\,p(B \mid A)$$

$$p(A \mid B) = \frac{p(A)\,p(B \mid A)}{p(B)}$$

GORNJI PRIMER AIDS TESTA

Bayes kaže:

$$P(\text{inf}\,|+) = \frac{P(+|\text{inf})P(\text{inf})}{P(+|\text{inf})P(\text{inf}) + P(+|\neg\text{inf})P(\neg\text{inf})}$$

"Evropski prior":

$$P(\text{inf}) = 1 - P(\neg\text{inf}) = 10^{-3}$$

Bayes računa:

$$P(\text{inf}\,|+) = \frac{1\cdot 10^{-3}}{1\cdot 10^{-3} + 0,002\cdot(1-10^{-3})} \approx 0,33$$

I pored pozitivnog ishoda na testu, i dalje imate razlog da se kladite 2:1 da je u pitanju greška testa!

ZAŠTO PREDSTAVLJATI STEPEN POVERENJA VEROVATNOĆAMA?

- Dobra statistika!
  - Konzistencija i greške u najgorem slučaju (*worst-case error*).
- "Holandska knjiga" + preživljavanje najsposobnijih
  - Ako su vaša verovanja u raskoraku sa zakonima verovatnoće, tada će vas u klađenju uvek pobediti neko čija su verovanja bliže tim zakonima.
- Daje nam teoriju postepenog učenja!
  - Uobičajena procedura za kombinovanje prethodnog (a priori) znanja sa novim iskustvima.

HIPOTEZE U BAJESOVSKOM ZAKLJUČIVANJU

- Hipoteze H se odnose na procese koji su mogli *generisati* podatke D
- Bajesovsko zaključivanje nam daje raspodelu na prostoru tih hipoteza, uz dato D
- $P(D|H)$ je verovatnoća da je D generisano procesom koji identifikuje H
- Hipoteze H su **međusobno isključive**: samo jedan proces je mogao generisati D

PRIMER: BELGIJSKI EVRO

- Belgijski novčić od 1€ bačen N = 250 puta dao je ishod “glava” X = 140 puta.
- “*It looks very suspicious to me*. If the coin were unbiased the chance of getting a result as extreme as that would be less than 7%”
  – Barry Blight, LSE (*Guardian*, 2002)

KLASIČNO TESTIRANJE HIPOTEZA

- Nulta hipoteza $H_0 : \theta = 0.5$ (“ispravan” novčić)
- U klasičnoj analizi ne moramo da specifikujemo alternativne hipoteze, ali kasnije ćemo koristiti $H_1 : \theta \neq 0.5$
- Potrebno nam je pravilo odlučivanja koje mapira podatke D na par prihvatamo / odbacujemo H_0.
- Definišimo skalarnu meru odstupanja od nulte hipoteze (“devijansu”) d(D) npr. χ^2.

P-VREDNOSTI

- Definišimo p-vrednost na kritičnom pragu τ kao

$$pval(\tau) = P(\{D : d(D) \geq \tau\}|H_0, N)$$

- Intuitivno, p-vrednost podataka je verovatnoća da se dobiju *najmanje podjednako ekstremni* rezultati ako važi H_0.
- Obično biramo τ tako da eventualno lažno odbacivanje H_0 bude ispod nivoa značaja $\alpha = 0.05$

$$P(R_\tau|H_0, N) \leq \alpha$$

P-VREDNOST ZA EVRO

- N = 250 ogleda, X=140 “glava”
- P-vrednost je “ispod 7%”

$$pval = P(X \geq 140|H0, N) + P(X \leq 110|H0, N) = 0.066$$

Pval=(1-binocdf(139,n,0.5)) + binocdf(110,n,0.5)

- Ako bi bilo N=250 i X=141, pval = 0.0497, tako da možemo odbaciti nultu hipotezu na nivou značaja 5%.
- Ovo ne znači da je $P(H_0|D)=0.07$!

BAJESOVSKA ANALIZA BELGIJSKOG EVRA

- Pretpostavimo $P(H_0) = P(H_1) = 0.5$
- Neka je $P(p) \sim p^{\alpha} (1-p)^{\alpha} = \text{Beta}(\alpha,\alpha)$
- Postavljanjem $\alpha = 1$ dobijamo uniformni (neinformativni) prior.

$$B = \frac{P(H_1|D)}{P(H_0|D)} = \frac{P(D|H_1)}{P(D|H_0)}\frac{P(H_1)}{P(H_0)}$$
$$= \frac{Z(\alpha_h + N_h, \alpha_t + N_t)}{Z(\alpha_h, \alpha_t)} \times \frac{1}{2}^N$$

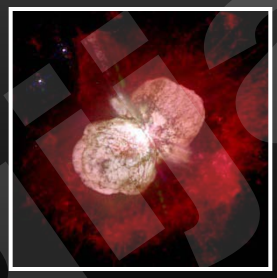
BAJESOVSKA ANALIZA (II)

- Ako α=1, $B = \frac{P(H_1|D)}{P(H_0|D)} = 0.45$ tako da je H_0 (“ispravan”) nešto verovatnije od H_1 (“neispravan”).
- Varirajući α preko velikog raspona, najbolje što možemo da dobijemo je B=1.9, što ne podržava hipotezu o neispravnosti novčića!
- Drugačiji priori daju slične rezultate.
- **Bajesovska analiza je u suprotnosti sa klasičnom analizom!**

PRINCIP MAKSIMALNE ENTROPIJE

- Jaynes (1957): najopravdaniji model je onaj koji maksimalizuje Šenonovu (informacionu) entropiju $H(\boldsymbol{p}) = - \Sigma\, p_i \log p_i$ za skup hipoteza konzistentnih sa zadatim skupom podataka.
- Informaciona entropija meri "neinformativnost" hipoteze; može varirati: 0 (potpuno informativna) $\leftrightarrow \log n$ (potpuno neinformativna).
- Okamova oštrica: "najjednostavnija" (= najmanje informativna!) hipoteza konzistentna sa svim podacima!
- Postoji nekoliko egzaktnih izvođenja koja su analogna izvođenju Meksvel-Bolcmanove raspodele u klasičnoj statističkoj mehanici.
- Ajnštajn: "Sve treba da bude najjednostavnije moguće, ali ne više od toga!"

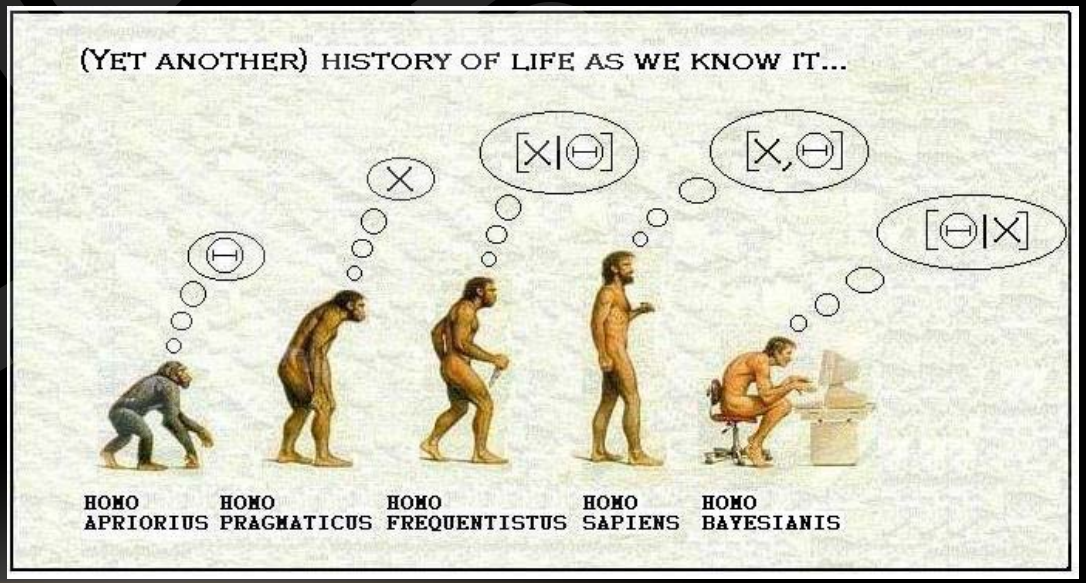
PROBLEM SA BAJESOM: "ARGUMENT SUDNJEG DANA"



- Brandon Carter oko 1990; John Leslie 1992; Richard Gott 1996
- Ogled sa dva ćupa
- Dva modela istorije čovečanstva?
- Ukupan broj ljudi koji su živeli do danas je $\sim 6 \times 10^{10}$.
- Kolika je verovatnoća da imamo tako nizak rang pod "optimističkom" hipotezom? Očigledno mala $\Rightarrow$ **Sudnji dan je blizu!**
- Postoje razna rešenja (recimo drugačija definicija individue), ali nijedno nije definitivno – problem i dalje otvoren!

MALO BAJESOVSKOG HUMORA

- “A Bayesian is one who, vaguely expecting a horse, and catching a glimpse of a donkey, strongly believes he has seen a mule.”



UMESTO ZAKLJUČKA...

- Verovatnoća je mnogo složenija stvar nego što se to čini!
- Bajesovski pristup verovatnoći je plodotvorniji od frekventističkog kad su prirodne nauke u pitanju.
- Bajes reprodukuje sam ljudski proces učenja, tj. **ažuriranje naših predstava o svetu**. Radimo to što češće!

PAR KORISNIH REFERENCI

- E. T. Jaynes. *Probability Theory: The Logic of Science* (Cambridge University Press, 2003).
- http://bayes.wustl.edu/
- G. D'Agostini, *Bayesian Reasoning in Data Analysis: A Critical Introduction* (World Scientific, 2003).
- H. Poincare, *Science and Hypothesis* (1905, Dover ed. 1952).
- C. Howson & P. Urbach, *Scientific Reasoning: the Bayesian Approach* (Open Court Publishing Company, 2005).